DOCUMENT RESUME

ED 367 707 TM 021 175

AUTHOR Thompson, Bruce
TITLE It Is Incorrect To Say "The Test Is Reliable": Bad Language Habits Can Contribute to Incorrect or Meaningless Research Conclusions.
PUB DATE Jan 94
NOTE 32p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 27-29, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Data Collection; Editing; Jargon; Language Patterns; Language Role; *Language Usage; Quality Control; *Reliability; Research Problems; *Research Reports; Scores; Semantics; Standards; Test Theory; Validity; *Writing for Publication; Written Language
IDENTIFIERS *Accuracy; *Meaningfulness

ABSTRACT

Researchers too frequently consider the reliability of the scores they analyze, and this may lead to incorrect conclusions. Practice in this regard may be negatively influenced by telegraphic habits of speech implying that tests possess reliability and other measurement characteristics. Styles of speaking in journal articles, in textbooks, and in professional standards and guidelines are explored. Two recommendations are offered. First, the statement "the test is reliable" should be recognized as being inappropriate, and professional standards and editorial guidelines should make this clear. Second, an important implication of the realization that reliability inures to data, rather than tests, is that reliability should generally be explored whenever data are collected. Three tables present language usage examples. An appendix lists 52 articles surveyed. (Contains 28 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 367 707

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BRUCE THOMPSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

It is Incorrect to Say "The Test is Reliable":

Bad Language Habits

Can Contribute to Incorrect or Meaningless Research Conclusions

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 27, 1994. The author acknowledges the thoughtful comments of Jerome T. Kapes on a previous version of this paper.

TM021175

BEST COPY AVAILABLE

ABSTRACT

Researchers too infrequently consider the reliability of the scores they analyze, and this may lead to incorrect conclusions. Practice in this regard may be negatively influenced by telegraphic habits of speech implying that tests possess reliability and other measurement characterics. Styles of speaking in journal articles, in textbooks, and in professional standards and guidelines, are explored. Suggestions for improved practice are presented.

Most of us, in both our daily lives and in our scholarship, are guided in our behavior by our paradigms. As defined by Gage (1963, p. 95), "Paradigms are models, patterns, or schemata. Paradigms are not the theories; they are rather ways of thinking or patterns for research." Tuthill and Ashton (1983, p. 7) explained that:

> A scientific paradigm can be thought of as a socially shared cognitive schema. Just as our cognitive schema provide us, as individuals, with a way of making sense of the world around us, a scientific paradigm provides a group of scientists with a way of collectively making sense of their scientific world.

But scholars usually do not consciously recognize the influence of their paradigms. As Lincoln and Guba (1985, pp. 19-20) noted:

> If it is difficult for a fish to understand water because it has spent all its life in it, so it is difficult for scientists... to understand what their basic axioms or assumptions might be and what impact those axioms and assumptions have upon everyday thinking and lifestyle.

Even though social scientists are usually unaware of paradigm influences, paradigms nevertheless are potent influences in that they tell us what we need to think about, and also the things <u>about which we need not think</u>. As Patton (1975, p. 9) suggested,

> Paradigms are normative, they tell the practitioner what to do without the necessity of long existential or epistemological consideration. But it is this aspect of a paradigm that constitutes both its strength and its weaknesses--its strength in that it makes action possible; its weakness in that the very reason for action is hidden in the unquestioned assumptions of the paradigm.

Although scholars are usually blind to the impacts of their paradigms, occasionally paradigm presumptions "leak out" in the language that scientists use. Conversely, the things we say conventionally, even when our jargon has become telegraphic shorthand, can subsequently come to be perceived by us as *literal* truth, and then unquestioned, within the context of our paradigms.

One common feature of contemporary scholarly language is the usage of the statement, "the test is reliable." The purpose of this essay is to argue that such language is both incorrect and deleterious in its affects on scholarly inquiry, particularly given the pernicious consequences that unconscious paradigmatic beliefs can exact.

The paper the nature of reliability is reviewed, and then the consequences of insufficiently considering reliability when conducting substantive research addressing basic and applied problems is considered. Next, language use in one prominent journal is reviewed, related language use in four prominent textbooks is reviewed, and then language use in profesional

standards and guidelines is considered. Finally, suggestions for improved practice are presented.

The Nature of Reliability

Too few researchers act on a conscious recognition that *reliability is a characteristic of scores or the data in hand.* Many authors present this view, but paradigm influences constrain some researchers from actively integrating this presumption into their actual analytic practice.

As Rowley (1976, p. 53, emphasis added) argued, "It needs to be established that an instrument itself is neither reliable nor unreliable.... A single instrument can produce *scores* which are reliable, and other *scores* which are unreliable." Similarly, Crocker and Algina (1986, p. 144, emphasis added) argued that, "...A test is not 'reliable' or 'unreliable.' Rather, reliability is a property of the *scores* on a test for a *particular* group of examinees."

In another widely respected text, Gronlund and Linn (1990, p. 78, emphasis in original) noted,

> Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself.... Thus, it is more appropriate to speak of the reliability of the "test scores" or of the "measurement" than of the "test" or the "instrument."

And Eason (1991, p. 84, emphasis added) argued that:

> Though some practitioners of the classical

> measurement paradigm [incorrectly] speak of reliability as a characteristic of tests, in fact reliability is a characteristic of *data*, albeit data generated on a given measure administered with a given protocol to given subjects on given occasions.

The subjects themselves impact the reliability of scores, and thus it becomes an oxymoron to speak of "the reliability of the test" without considering to whom the test was administered, or other facets of the measurement protocol. Reliability is driven by variance--typically, greater scores variance leads to greater score reliability, and so more *heterogeneous* samples often lead to more *variable* scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogenous or to more homogeneous sets of subjects, will yield scores with differing reliability. As Dawes (1987, p. 486) observed, "...Because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population--an obvious but sometimes overlooked point."

Our shorthand ways of speaking (e.g., language saying "the test is reliable") can itself cause confusion and lead to bad practice. As Pedhazur and Schmelkin (1991, p. 82, emphasis in original) observed, "Statements about *the* reliability of a measure are... inappropriate and potentially misleading." These telegraphic ways of speaking are not inherently problematic, but they often later become so when we come unconsciously to ascribe literal truth to our shorthand, rather than recognizing that our

jargon is sometimes telegraphic and is not literally true. As noted elsewhere:

> This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. Thompson (1992, p. 436)

The Important Impacts of Reliability on Substantive Research

In one book exploring the intimate linkages between measurement error variance and our attributions about the origins of variance in our substantive basic or applied research research, Pedhazur and Schmelkin (1991) noted,

> Measurement error is the Achilles' heel of sociobehavioral research. Although most programs in sociobehavioral sciences, especially doctoral programs, require a modicum of exposure to statistics and research design, few seem to require the same where measurement is concerned. Thus, many students get the impression that no special competencies are necessary for the development and use of measures... (pp. 2-3)

Therefore, it should not be surprising that studies of research reports in journals indicate insufficient attention to the impacts of measurement integrity on the integrity of substantive research conclusions. For example, with respect to the American

Educational Research Journal, Willson (1980) reported that:

> ...Only 37% of the AERJ studies explicitly reported reliability coefficients for the data analyzed. Another 18% reported only indirectly through reference to earlier research.... That reliability... is unreported in almost half the published research is... inexcusable at this late date...." (pp. 8-9)

A more recent "perusal of contemporary psychology journals demonstrates that quantitative reports of scale reliability and validity estimates are often missing or incomplete" (Meier & Davis, 1990, p. 113); and that "the majority [95%, 85% and 60%] of the scales described in the [three Journal of Counseling Psychology] JCP volumes [1967, 1977 and 1987] were not accompanied by reports of psychometric properties" (p. 115). The situation is apparently roughly equivalent as regards dissertation research (Thompson, 1988).

This state of affairs is surprising, given two related trends within the literature. First, since the influential articles by Cohen (1968) and Knapp (1978) appeared, more researchers have recognized that all parametric statistical analyses are correlational (Thompson, 1991), and that substantive variance-accounted-for effect sizes expressed as r^2 analogs can be interpreted in all studies. Second, the importance of interpreting effect sizes as against statistical significance tests has been increasingly recognized (e.g., Thompson, 1993), as reflected, for

example, in a recent cascade of articles within the American Psychologist (cf. Cohen, 1990; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989).

Nevertheless, too few researchers act on the premise that score reliability establishes a ceiling for substantive effect sizes. These impacts can be readily illustrated in a concrete example using the bivariate correlation as an heuristic.

It has been recognized in textbooks dating back to the 1950s, and in more recent books as well (e.g., Pedhazur & Schmelkin, 1991, p. 114), that a correlation coefficient "corrected" for attenuation due to measurement error ($\hat{r}_{XY}$) can be estimated as:

$$\hat{r}_{XY} = r_{XY} / [(r_{XX} * r_{YY})^{.5}],$$

where r_{XY} is the calculated bivariate relationship between scores on variables X and Y, and r_{XX} and r_{YY} are respectively the reliability coefficients for scores on X and Y. This algorythm can be re-expressed in the more familiar metric of common variance, as is often done in popular variance-accounted-for effect size statistics (e.g., r^2, R^2, eta^2, $omega^2$):

$$\hat{r}_{XY}^2 = r_{XY}^2 / (r_{XX} * r_{YY})$$

Through algebraic manipulation, the detectable effect size, given knowledge of "true" relationship, $\hat{r}_{XY}^2$, and the reliabilities of the two sets of scores, is:

$$r_{XY}^2 = \hat{r}_{XY}^2 * (r_{XX} * r_{YY})$$

Even if the "true" relationship between perfectly reliable measures of X and Y was perfect, i.e., $\hat{r}_{XY}^2 = 1.0$, the detectable effect in any study can *never exceed the product of the reliability*

coefficients for the two sets of scores:

$$r_{XY}^2 = 1 * (r_{XX} * r_{YY})$$

For example, even when $\hat{r}_{XY}^2 = 1.0$, if both sets of scores have reliability coefficients of .7, the detectable effect cannot exceed .49. Clearly, measurement error prospectively impacts the effect size that we can obtain in a planned study and also should be retrospectively considered when interpreting calculated effects once the study has been done.

The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. We may conduct studies that could not possibly yield noteworthy effect sizes. Or we may not accurately interpret our results if we do not consider the reliability of the scores we are actually analyzing.

These practices may be caused by misperceptions that tests can be reliable or valid. These misperceptions themselves may be caused, or at least reinforced, by the use of telegraphic language that comes to be unconsciously believed as literal truth, and then unconsciously incorporated into paradigms for behavior.

Language Use in A Prominent Measurement Journal

Logically, if the language used by the best experts to describe measurement integrity was telegraphic or inappropriate, then, *a fortiorari*, appropriate language use and thinking by others regarding score reliability would be even less likely. One empirical snapshot of contemporary language practice was derived for the present paper by reviewing all the articles in the

measurement integrity studies section of Educational and Psychological Measurement (EPM). EPM is a journal that was started some 50 years ago with Frederick Kuder as Founding Editor, and until very recently, also as the journal's owner. Kuder, of course, is widely known for his contributions to reliability theory through the various "KR" formulas.

The 1992 volume of EPM contained 64 articles in the journal's measurement integrity section. Eleven of these articles did not directly deal with measurement characteristics issues. One of the remaining 53 articles involved the present author as a coauthor, and did not involve the language use issues described here. Table 1 presents illustrative quotations from the remaining 52 articles. The tabled quotations, even in a respected forum presumably involving measurement experts as authors and reviewers, reflect a pattern of language usage regarding measurement characteristics that is at best telegraphic in style.

INSERT TABLE 1 ABOUT HERE.

Language Use in Four Prominent Measurement Texts

Four well-known measurement textbooks (Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Sax, 1989; Thorndike, Cunningham, Thorndike, & Hagen, 1991) were also surveyed to garner an impression of language use as regards score reliability. Table 2 presents illustrative quotations from these works. Even respected texts being published in as late as 6th editions reflect language usage that is at best inconsistent, telegraphic, or incorrect.

INSERT TABLE 2 ABOUT HERE.

One set of authors, for example, presents an oxymoron in which it is asserted that (a) the sample impacts reliability but that (b) somehow over different samples still "the test is reliable". These authors note, "A third factor influencing the estimated *reliability of a test* is group homogeneity" (Mehrens & Lehmann, 1991, p. 259, emphasis added).

Language Use in Professional Standards and Guidelines

The language in professional journals and textbooks has both infuelnced and been influenced by the language use in professional standards and guidelines. For example, Meier and Davis (1990, p. 113) suggested that so few authors may test or even discuss the reliability of their scores partially as

> ...the result of a lack of explicit guidelines for the reporting of scale information. For example, the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983) makes no specific recommendations in regard to the reporting of scales' psychometric properties.

Table 3 reports related language use in two fairly recent sets of professional standards (APA/AERA/NCME, 1985; Joint Committee, in press). For example, the APA/AERA/NCME (1985) test standards emphasize that, "Because there are many ways of estimating reliability, each influenced by different sources of measurement error, it is unacceptable to say simply, 'The reliability of test

X is .90'" (p. 21). Yet, on the same page, these standards speak of "the *reliability of a highly speeded test*" (APA/AERA/NCME, 1985, p. 21, emphasis added).

INSERT TABLE 3 ABOUT HERE.

Conclusions

Based on these considerations, two recommendations are offered. First, the language of saying "the test is reliable" should be recognized as being inappropriate, and professional standards and editorial guidelines should make forcefully this clear. Instead, authors should be encouraged to say, "the scores in our study had a classical theory test-retest reliability coefficient of X," or "based on generalizability theory analysis, the scores in our study had a phi coefficient of X."

It will not be sufficient to say in our standards that, "Because there are many ways of estimating reliability, each influenced by different sources of measurement error, it is unacceptable to say simply, 'The reliability of test X is .90'" (APA/AERA/NCME, 1985, p. 21). Rather, such language usage should be declared inappropropriate because the language is, on its face, untrue. And the consequences of believing untrue shorthands should be noted within our professional standards.

Of course, the illustrations of language use presented in Tables 1 through 3 suggest that changing our habits of speech will be a daunting task. But, as Lachman (1993) noted, "Language habits are difficult to change. Sometimes, however, it is appropriate and

desirable to change them" (p. 1093).

Second, as suggested elsewhere,

> One important implication of the realization that reliability inures to data (rather than tests) is that reliability should generally be explored whenever data are collected. And we always need to thoughtfully and explicitly explore whether the data in hand were collected on a sample similar to the samples used in previous reliability studies with a given measure. (Thompson, 1992, p. 436)

Such practices would provide better models for behavior, would provide more information in the literature about the data from our measures, and would themselves challenge paradigmatic assumptions that "the test is [or can be] reliable."

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (APA/AERA/NCME). (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Dawes, R.V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Gage, N.L. (1963). Paradigms for research on teaching. In N.L. Gage (Ed.), Handbook of research on teaching (pp. 94-141). Chicago: Rand McNally.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Joint Committee on Standards for Educational Evaluation. (in press). The program evaluation standards: How to assess

evaluations of educational programs. Newbury Park, CA SAGE.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Lachman, S.J. (1993). Statistically significant differences or probable nonchance difference. American Psychologist, 48, 1093.

Lincoln, Y.S., & Guba, E.G. (1985). Naturalistic inquiry. Newbury Park, CA: SAGE.

Meier, S.T., & Davis, S.R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. Journal of Counseling Psychology, 37, 113-115.

Mehrens, W.A., & Lehmann, I.J. (1991). Measurement and evaluation in education and psychology (4th ed.). Fort Worth: Holt, Rinehart and Winston.

Patton, M.Q. (1975). Alternative evaluation research paradigm. Grand Forks: University of North Dakota Press.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science.

American Psychologist, 44, 1276-1284.

Rowley, G.L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadsworth.

Thompson, B. (1988, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (Ed.). (1993). Special issue on statistical significance testing, with comments from various journal editors, Journal of Experimental Education, 61(4).

Thorndike, R.M., Cunningham, G.K., Thorndike, R.L., & Hagen, E.P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: MacMillan.

Tuthill, D., & Ashton, P. (1983). Improving educational research through the development of educational paradigms. Educational Researcher, 12(10), 6-14.

Willson, V.L. (1980). Research techniques in AERJ articles: 1969 to

1978. Educational Researcher, 9(6), 5-10.

Table 1
Illustrative Journal Quotations Illustrating Telegraphic Language
(Emphases Added to All Quotations)

Examples of Telegraphic Misspeaking

"The Speed of Thinking Test appears to provide a *measure* of cognitive speed that *is* sufficiently *reliable* and valid..." (Carver, 1992, p. 132)

"...a major shortcoming of research in this domain has been the lack of *a reliable and valid measure*..." (Schriesheim, Neider, Scandura & Tepper, 1992, p. 136)

"The internal consistency *reliability for the scales* and subscales were calculated using Cronbach's alpha." (Caruso, 1992, p. 156)

"...the *MBI possesses* an acceptable level of *reliability*..." (Abu-Hilal, M.M., & Salameh 1992, p. 168).

"...the internal consistency *reliabilities* (coefficient alpha) *of the new scales* were computed..." (Romero, Tepper & Tetrault, 1992, p. 176)

"The results of this study suggest that the *scale* developed here *is* highly *reliable*..." (Murphy & Thorton, 1992, p. 199).

"...the *SWMSS possessed strong reliability*, and convergent and discriminant validity..." (Vandenberg & Scarpello, 1992, p. 204)

"...Cronbach's alpha showed that the overall *reliability of* the 20-item *scale* was..." (Chow & Winzer, 1992, p. 227)

"Evidence on the *reliability*, stability, and validity *of the NEO-PI* has been reviewed..." (McCrae & Costa, 1992, p. 232)

"The results of the statistical analyses indicate that the Student Religiosity *Questionnaire provides a reliable measure*..." (Katz & Schmida, 1992, p. 355)

"The concurrent *validity of the MTA scale* was supported..." (d'Ailly & Bergering, 1992, p. 370)

"...a lack of predictive *validity of this subtest* in medical education." (Glaser, Hojat, Veloski, Blacklow & Goepp, 1992, p. 405)

"*Reliability of the* 20-item *scale* was determined using coefficient alpha..." (Smither & Houston, 1992, p. 414)

"The *instrument* used to measure comprehension monitoring ability was found *to have substantial reliability*..." (Otero, Campanario & Hopkins, 1992, p. 428)

"Reliability data... show adequate to high *alpha coefficients to [sic] each subscale...*" (Thornburg, Ispa, Adams & Lee, 1992, p. 432)

"...most of these developing *abilities* were also the ones that *had* high 4-year test-retest *reliabilities...*" (Dawis, Goldman & Sung, 1992, p. 464)

"One approach to determining construct *validity of a test* is to examine item content..." (Wooley & Hakstian, 1992, p. 476)

"...the *items are* more *valid* for men than for women." (Novy, 1992, p. 494)

"...this *measure is not reliable...*" (Rentsch & Heffner, 1992, p. 646)

"...*a reliable... measure* of computer attitudes among professional nurses." (Coover & Delcourt, 1992, p. 654)

"...establish the construct *validity of a psychometric instrument* for assessing beliefs..." (Silvernail, 1992, p. 667)

"The *validity of* such *instruments...*" (Austin, 1992, p. 669)

"After examining four *inventories*, Biaggio (1980) questioned their construct validity..., *their poor reliability*, and limited predictive validity." (Kroner, Reddon & Serin, 1992, p. 688)

"...the shorter *scales are* a little *less reliable* than the longer scales..." (Francis & Katz, 1992, p. 697)

"...the comparative *validity of the two measures...*" (Goldstein & Bokoros, 1992, p. 707)

"...*reliabilities of the item sets* were moderate..." (Beyler & Schmeck, 1992, p. 713)

"If the *subtests* weighted in this process *were not valid...*" (Earles & Ree, 1992, pp. 722)

"...the *reliability* and validity... *of* two American-developed *instruments...*" (Watkins & Gerong, 1992, p. 728)

"The obtained estimates of internal-consistency *reliability for the* Revised Maslach Burnout *Scale* was .82..." (Gryskiewicz & Buttner, 1992, p. 749)

"The internal-consistency *reliability* coefficient (coefficient alpha) *for the scale* was 0.90.... It would also appear to be a valid instrument..." (Pretorius & Norman, 1992, pp. 936-937)

"...the construct *validity of the original LSI*..." Geiger, Boyle, & Pinto, 1992, p. 758)

"...it is important that *a valid measure* be found..." (Gold, Roth, Wright, Michael & Chen, 1992, p. 762)

"...whether or not the predictive *validity of the Leniency Scale* would be affected..." (Highhouse, 1992, p. 785)

"These achievement *tests have reliability* estimates greater than .92." (Marjoribanks, 1992, p. 947)

"...the lack of valid and *reliable instruments*..." (Short & Rinehart, 1992, p. 953)

"Once the *reliability of the Anxiety Scale* had been established..." (Sánchez-Herrero & Sánchez, 1992, p. 964)

"...the Cultural Literacy *Test is very reliable*..." (Pentony, 1992, p. 970)

"...question the *validity of the instrument*..." (Ayers & Quattlebaum, 1992, p. 973)

"Both of these scales... have evidence supporting *their reliability* and validity..." (Schriesheim, Scandura, Eisenbach & Neider, 1992, p. 985)

"With respect to the *reliability of the scale*, results from this study revealed that the internal consistency of all subscales was adequate..." (Vallerand, Pelletier, Blais, Brière, Senécal & Vallières, 1992, p. 1015)

"...the *test has* demonstrated high *reliability*..." (Goldberg & Alliger, 1992, p. 1022)

"The two *halves of the SCT have* internal-consistency estimates of *reliabilities* greater than .80. (Novy & Francis, 1992, p. 1038)

"*Cronbach's alpha for the SL-ASIA* was found to be .91..." (Suinn, Ahuna & Khoo, 1992, p. 1043)

"...the *SAT has* even less incremental *validity* than their results suggest..." (Baron & Norman, 1992, p. 1054)

Anthropometric Attribution to Tests Being Actors

"The three satisfaction *instruments* in the study *displayed* reasonable levels of internal consistency *reliability*." (Rentsch & Steel, 1992, p. 360)

"...this shortened evaluation *instrument demonstrates* very high *reliability*..." (Fernández & Mateo, 1992, p. 679)

"The obtained factor solutions and resulting reliability coefficients for the CAS, CARS, and CSE suggest that each *instrument exhibits* construct validity and *reliability*." (Harrison & Rainer, 1992, p. 744)

Measurement Characteristics Ascribed to Model/Theory

"Further studies are needed to shed light on the *validity of* the Crites *model*..." (Westbrook & Sanford, 1992, p. 351)

Inconsistent Use of Language

"The reliability coefficients for the creativity composites [i.e., scores] were... The reliability coefficients for the Intelligence ratings were..." (Runco & Mraz, 1992, p. 217)

versus

"The new *scoring technique... has* demonstrated *reliability*." (Runco & Mraz, 1992, p. 219)

"Internal-consistency estimates of reliability for the total score across the grade levels is adequate..." (Hagborg & Wachman, 1992, p. 438)

versus

"...the *validity of the instrument* was supported..." (Hagborg & Wachman, 1992, p. 438)

"The reliability and validity of obtained raw scores were virtually unaffected..." (Simpson & Halpin, 1992, p. 468)

versus

"...no accompanying loss in *reliability* or validity *of the test*..." (Simpson & Halpin, 1992, p. 468)

"The K-BIT manual reports an internal consistency coefficient of .92 for the total sample and test-retest reliability coefficients greater than .90 for each age group." (Prewitt, 1992, p. 979)

versus

"...the K-BIT should have evidence supporting *its concurrent validity*..." (Prewitt, 1992, p. 977)

Note. The reference list of these and other EPM articles surveyed is available from the author upon request.

Table 2
Illustrative Book Quotations Illustrating Language Use

(Thorndike, Cunningham, Thorndike, & Hagen, 1991)

"...The larger a sample of a person's behavior we have, the more reliable *the measure* will be." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 100, emphasis added)

"...*the test* with the higher reliability coefficient..." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 104, emphasis added)

"...we prefer the more *reliable test*." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 105, emphasis added)

"...to evaluate the reliability *of a test*..." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 118, emphasis added)

"...the correct reliability *for any instrument*." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 120, emphasis added)

"How reliable *a test* must be..." (Thorndike, Cunningham, Thorndike, & Hagen, 1991, p. 120, emphasis added)

(Gronlund & Linn, 1990)

"Any particular *instrument* may have a number of different reliabilities..." (Gronlund & Linn, 1990, p. 78, emphasis added)

"...constructing more *reliable classroom tests*." (Gronlund & Linn, 1990, p. 93, emphasis added)

"...the reliability of their own classroom *tests*." (Gronlund & Linn, 1990, p. 93, emphasis added)

"In general, the longer the test is, the higher *its reliability* will be." (Gronlund & Linn, 1990, p. 93, emphasis added)

"...effect on the reliability *of the measures* obtained..." (Gronlund & Linn, 1990, p. 97, emphasis added)

"...classroom *tests* of questionable reliability..." (Gronlund & Linn, 1990, p. 100, emphasis added)

versus

"...for estimating the reliability *of test scores*." (Gronlund & Linn, 1990, p. 83, emphasis added)

"...in estimating the reliability *of test scores*..." (Gronlund & Linn, 1990, p. 86, emphasis added)

"...provide more reliable *results*..." (Gronlund & Linn, 1990, p. 93, emphasis added)

"...the reliability of the test *results*..." (Gronlund & Linn, 1990, p. 97, emphasis added)

"...the reliability of our crtiterion-referenced *interpretations* with these tests." (Gronlund & Linn, 1990, p. 100, emphasis added)

"In interpreting and using reliability information, it is important to remember that reliability estimates refer to the *results* of measurement..." (Gronlund & Linn, 1990, p. 103, emphasis in original)

(Mehrens & Lehmann, 1991)

"...No *measure* is perfectly reliable." (Mehrens & Lehmann, 1991, p. 249, emphasis added)

"...should result in a reasonably reliable *test*." (Mehrens &

Lehmann, 1991, p. 249, emphasis added)
"...estimate the reliability of their own *instruments*..." (Mehrens & Lehmann, 1991, p. 249, emphasis added)
"In physical measurement we can ordinarily obtain very reliable *measures*." (Mehrens & Lehmann, 1991, p. 249, emphasis added)
"...an estimate of the *reliability* (or interindividual variability) of the measure." (Mehrens & Lehmann, 1991, p. 250, emphasis in original)
"...the more consistent (reliable) the *measurement*." (Mehrens & Lehmann, 1991, p. 250, emphasis added)
"...estimates of the reliability of their classroom *tests*." (Mehrens & Lehmann, 1991, p. 256, emphasis added)
"...to estimate what the reliability of a *test* would be..." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"...if a *test* has an original reliability..." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"Just as adding equivalent items makes a test *score* more reliable, so deleting equivalent items makes a *test* less reliable." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"...a *test* with low reliability..." (Mehrens & Lehmann, 1991, p. 263, emphasis added)
"...complained about standardized *tests* because they lack perfect reliability." (Mehrens & Lehmann, 1991, p. 264, emphasis added)

versus

"Technically speaking, *data* should be reliable, and the inferences we draw from the data should be valid." (Mehrens & Lehmann, 1991, p. 248, emphasis added)
"...the reliability of a set of *scores*." (Mehrens & Lehmann, 1991, p. 248, emphasis added)
"...the reliability of the sum (or average) of the two readers' *scores*..." (Mehrens & Lehmann, 1991, p. 257, emphasis added)
"...longer tests give more reliable *scores*." (Mehrens & Lehmann, 1991, p. 258, emphasis added)
"The reliability of the *data*..." (Mehrens & Lehmann, 1991, p. 262, emphasis added)
"...the *data* should be fairly reliable..." (Mehrens & Lehmann, 1991, p. 262, emphasis added)
"...the reliability of the *test*..." (Mehrens & Lehmann, 1991, p. 262, emphasis added)
"...the reliability of the *scores* is of more concern..." (Mehrens & Lehmann, 1991, p. 263, emphasis added)
"...the *scores* should be more reliable..." (Mehrens & Lehmann, 1991, p. 263, emphasis added)
"...consider the quality of the *data*. Reliability is one of the more important qualities." (Mehrens & Lehmann, 1991, p. 264, emphasis added)

(Sax, 1989)

"Unreliable *tests* measure the effects of chance..." (Sax, 1989, p. 259, emphasis added)
"A *test* with low reliability..." (Sax, 1989, p. 259, emphasis added)

"...consideration of the reliability *of measurements*. Unreliable *tests* are no better..." (Sax, 1989, p. 259, emphasis added)
'reliability of a test' should always be interpreted to mean the 'reliability of measurements or observations derived from a test.'" (Sax, 1989, pp. 263-264, emphasis in original)
"Parallel [test] *forms* are never perfectly correlated or reliable." (Sax, 1989, p. 264, emphasis added)

versus

"...It is more accurate to talk about the reliability of *measurements* (data, scores, and observations) than the reliability of *tests* (questions, items, and other tasks). Any reference to the +"...the reliability of *measurements*..." (Sax, 1989, p. 273, emphasis added)
"...total *scores* usually have higher reliabilities." (Sax, 1989, p. 275, emphasis added)

Table 3
Language Usage in Professional Standards

(Joint Committee on Standard for Educational Evaluation, in press)

[A common error is] "[f]ailing to take into account the fact that the *reliability of* the *scores* provided by an instrument or procedure may fluctuate depending on how, when, and to whom the instrument or procedure is administered." (Joint Committee on Standard for Educational Evaluation, in press, emphasis added)

"A generic term, *reliability* refers to the degree of consistency *of the information* obtained from an information gathering process." (Joint Committee on Standard for Educational Evaluation, in press)

"Whenever possible, evaluators should choose information gathering procedures that have, in the past, yielded *data* and information *with acceptable reliability* for their intended uses; however, the generalizability of previous favorable reliability results may not be simply assumed. Reliability information should be collected that is directly relevant to the groups and ways in which the information gathering procedures will be used in the evaluation." (Joint Committee on Standard for Educational Evaluation, in press)

(APA/AERA/NCME, 1985)

"*Reliability* refers to the degree to which test *scores* are free from errors of measurement." (APA/AERA/NCME, 1985, p. 19)

"Measurement errors reduce the *reliability* (and therefore the generalizability) *of the score* obtained for a person..." (APA/AERA/NCME, 1985, p. 19, emphasis added)

"But *scores* representing differences between scores obtained from two tests or from repeated administrations of the same test... are generally *less reliable* than either of the parts." (APA/AERA/NCME, 1985, p. 20, emphasis added)

APPENDIX A

List of Volume 52 EPM Articles (n=53+11=64) Surveyed

Studies of (n=53) Measurement Characteristics

Abu-Hilal, M.M., & Salameh, K.M. (1992). Validity and reliability of the Maslach Burnout Inventory for a sample of non-western teachers. Educational and Psychological Measurement, 52(1), 161-169.

Austin, J.S. (1992). The detection of fake good and fake bad on the MMPI-2. Educational and Psychological Measurement, 52(3), 669-674.

Ayers, J.B., & Quattlebaum, R.F. (1992). TOEFL performance and success in a masters program in engineering. Educational and Psychological Measurement, 52(4), 973-975.

Baron, J., & Norman, M.F. (1992). SATs, achievement tests, and high-school class rank as predictors of college performance. Educational and Psychological Measurement, 52(4), 1047-1055.

Beyler, J., & Schmeck, R.R. (1992). Assessment of individual differences in preferences for holistic-analytic strategies: Evaluation of some commonly available instruments. Educational and Psychological Measurement, 52(3), 709-719.

Burrell, B., Thompson, B., & Sexton, D. (1992). The measurement integrity of data collected using the Child Abuse Potential Inventory. Educational and Psychological Measurement, 52, 993-1001.

Caruso, G.L. (1992). The development of three scales to measure the supportiveness of relationships between parents and child care providers. Educational and Psychological Measurement, 52(1), 149-160.

Carver, R.P. (1992). Reliability and validity of the speed of thinking test. Educational and Psychological Measurement, 52(1), 125-134.

Chow, P., & Winzer, M.M. (1992). Reliability and validity of a scale measuring attitudes toward mainstreaming. Educational and Psychological Measurement, 52(1), 223-228.

Coover, D., & Delcourt, M.A.B. (1992). Construct and criterion-related validity of the Adult-Attitudes Toward Computers Survey. Educational and Psychological Measurement, 52(3), 653-661.

d'Ailly, H., & Bergering, A.J. (1992). Mathematics anxiety and mathematics avoidance behavior: A validation study of two MARS factor-derived scales. Educational and Psychological Measurement, 52(2), 369-377.

Dawis, R.V., Goldman, S.H., & Sung, Y.H. (1992). Stability and change in abilities for a sample of young adults. Educational and Psychological Measurement, 52(2), 457-465.

Earles, J.A., & Ree, M.J. (1992). The predictive validity of the ASVAB for training grades. Educational and Psychological Measurement, 52(3), 721-725.

Fernández, J., & Mateo, M.A. (1992). Student evaluation of university teaching quality: Analysis of a questionnaire for a sample of university students in Spain. Educational and

Psychological Measurement, 52(3), 675-686.
Francis, L.J., & Katz, Y. (1992). The comparability of the short form EPQ-R indices of extraversion, neuroticism, and the lie scale with the EPQ for a sample of 190 student teachers in Israel. Educational and Psychological Measurement, 52(3), 695-700.
Geiger, M.A., Boyle, E.J., & Pinto, J. (1992). A factor analysis of Kolb's Revised Learning Style Inventory. Educational and Psychological Measurement, 52(3), 753-759.
Glaser, K., Hojat, M., Veloski, J.J., Blacklow, R.S., & Goepp, C.E. (1992). Science, verbal, or quantitative skills: Which is the most important predictor of physician competence? Educational and Psychological Measurement, 52(2), 395-406.
Gold, Y., Roth, R.A., Wright, C.R., Michael, W.B., & Chen, C. (1992). The factorial validity of a teacher burnout measure (educators survey) administered to a sample of beginning teachers in elementary and secondary schools in California. Educational and Psychological Measurement, 52(3), 761-768.
Goldberg, E.L., & Alliger, G.M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. Educational and Psychological Measurement, 52(4), 1019-1027.
Goldstein, M.B., & Bokoros, M.A. (1992). Tilting at windmills: Comparing the Learning Style Inventory and the Learning Style Questionnaire. Educational and Psychological Measurement, 52(3), 701-708.
Gryskiewicz, N., & Buttner, E.H. (1992). Testing the robustness of the progressive phase burnout model for a sample of entrepreneurs. Educational and Psychological Measurement, 52(3), 747-751.
Hagborg, W.J., & Wachman, E.M. (1992). The Arlin Test of Formal Reasoning and the identification of accelerated mathematics students. Educational and Psychological Measurement, 52(2), 437-442.
Harrison, A.W., & Rainer, R.K., Jr. (1992). An examination of the factor structures and concurrent validities for the computer attitude scale, the computer anxiety rating scale, and the computer self-efficacy scale. Educational and Psychological Measurement, 52(3), 735-745.
Highhouse, S. (1992). The Leniency Scale: Is it really independent of ratee behavior? Educational and Psychological Measurement, 52(3), 781-786.
Katz, Y.J., & Schmida, M. (1992). Validation of the Student Religiosity Questionnaire. Educational and Psychological Measurement, 52(2), 353-356.
Kroner, D.G., Reddon, J.R., & Serin, R.C. (1992). The Multidimensional Anger Inventory: Reliability and factor structure in an inmate sample. Educational and Psychological Measurement, 52(3), 687-693.
McCrae, R.R., & Costa, P.T., Jr. (1992). Discriminant validity of NEO-PIR facet scales. Educational and Psychological Measurement, 52(1), 229-237.

Marjoribanks, K. (1992). The predictive validity of an attitudes-to-school scale in relation to children's academic achievement. Educational and Psychological Measurement, 52(4), 945-949.

Murphy, K.R., & Thorton, G.C., III. (1992). Development and validation of a measure of attitudes toward employee drug testing. Educational and Psychological Measurement, 52(1), 189-201.

Novy, D.M. (1992). Gender comparability of Forms 81 of the Washington University Sentence Completion Test. Educational and Psychological Measurement, 52(2), 491-497.

Novy, D.M., & Francis, D.J. (1992). Psychometric properties of the Washington University Sentence Completion Test. Educational and Psychological Measurement, 52(4), 1029-1039.

Otero, J., Campanario, J.M., & Hopkins, K.D. (1992). The relationship between academic achievement and metacognitive comprehension monitoring ability of Spanish secondary school students. Educational and Psychological Measurement, 52(2), 419-430.

Pentony, J.F. (1992). Cultural literacy: A concurrent validation. Educational and Psychological Measurement, 52(4), 967-972.

Pretorius, T.B., & Norman, A.M. (1992). Psychometric data on the statistics anxiety scale for a sample of South African students. Educational and Psychological Measurement, 52(4), 933-937.

Prewitt, P.N. (1992). The relationship between the Kaufman Brief Intelligence Test (K-BIT) and the WISC-R with incarcerated juvenille delinquents. Educational and Psychological Measurement, 52(4), 977-982.

Rentsch, J.R., & Heffner, T.S. (1992). Measuring self-esteem: Validation of a new scoring technique for "Who am I?" responses. Educational and Psychological Measurement, 52(3), 641-651.

Rentsch, J.R., & Steel, R.P. (1992). Construct and concurrent validation of the Andrews and Withey job satisfaction questionnaire. Educational and Psychological Measurement, 52(2), 357-367.

Romero, J.E., Tepper, B.J., & Tetrault, L.A. (1992). Development and validation of new scales to measure Kolb's (1985) learning style dimensions. Educational and Psychological Measurement, 52(1), 171-180.

Runco, M.A., & Mraz, W. (1992). Scoring divergent thinking tests using total ideational output and a creativity index. Educational and Psychological Measurement, 52(1), 213-221.

Sánchez-Herrero, S.A., & Sánchez, M.D.P. (1992). The predictive validity of an instrument designed to measure student anxiety in learning a foreign language. Educational and Psychological Measurement, 52(4), 961-966.

Schriesheim, C.A., Neider, L.L., Scandura, T.A., & Tepper, B.J. (1992). Development and preliminary validation of a new scale (LMX-6) to measure leader-member exchange in organizations. Educational and Psychological Measurement, 52(1), 135-147.

Schriesheim, C.A., Scandura, T.A., Eisenbach, R.J., & Neider, L.L. (1992). Validation of a new Leader-Member Exchange Scale (LMX-6) using heirarchically-nested maximum likelihood confirmatory factor analysis. Educational and Psychological Measurement, 52(4), 983-992).

Short, P.M., & Rinehart, J.S. (1992). School Participant Empowerment Scale: Assessment of level of empowerment within the school environment. Educational and Psychological Measurement, 52(4), 951-960.

Silvernail, D.L. (1992). The development and factor structure of the Educational Beliefs Questionnaire. Educational and Psychological Measurement, 52(3), 663-667.

Simpson, R.G., & Halpin, G. (1992). Psychometric effects of altering the ceiling criterion on the Passage Comprehension test of the Woodcock Reading Mastery Tests-Revised. Educational and Psychological Measurement, 52(2), 467-473.

Smither, R.D., & Houston, J.M. (1992). The nature of competitiveness: The development and validation of the Competitiveness Index. Educational and Psychological Measurement, 52(2), 407-418.

Suinn, R.M., Ahuna, C., & Khoo, G. (1992). The Suinn-Lew Asian Self-Identity Acculturation Scale: Concurrent and factorial validation. Educational and Psychological Measurement, 52(4), 1041-1046.

Thornburg, K.R., Ispa, J.M., Adams, N.A., & Lee, B.S. (1992). Testing the simplex assumption underlying the Erickson Psychosocial Stage Inventory. Educational and Psychological Measurement, 52(2), 431-436.

Vallerand, R.J., Pelletier, L.G., Blais, M.R., Brière, Senécal, C., & Vallières, E.F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. Educational and Psychological Measurement, 52(4), 1003-1017.

Vandenberg, R.J., & Scarpello, V. (1992). Multitrait-multimethod validation of the Satisfaction with My Supervisor Scale. Educational and Psychological Measurement, 52(1), 203-212.

Watkins, D., & Gerong, A. (1992). Evaluating undergraduate college teaching: A Filipino investigation. Educational and Psychological Measurement, 52(3), 727-734.

Westbrook, B.W., & Sanford, E.E. (1992). The relationship between career choice attitudes and career choice competencies of black 9th-grade pupils. Educational and Psychological Measurement, 52(2), 347-351.

Wooley, R.M., & Hakstian, A.R. (1992). An examination of the construct validity of personality-based and overt measures of integrity. Educational and Psychological Measurement, 52(2), 475-489.

Studies (n=11) NOT Investigating Measurement Characteristics

Brown, N.W., & Cross, E.J., Jr. (1992). A comparison of personality characteristics for entering freshmen, persistors, and norm groups in engineering. Educational and Psychological Measurement, 52(4), 939-944.

Farley, M.J., & Elmore, P.B. (1992). The relationship of reading comprehension to critical thinking skills, cognitive ability, and vocabulary for a sample of underachieving college freshmen. Educational and Psychological Measurement, 52(4), 921-931.

Fraboni, M., & Saltstone, R. (1992). The WAIS-R number-of-factors quandry: A cluster analytic approach to construct validation. Educational and Psychological Measurement, 52(3), 603-613.

Greenblatt, R.L., Mozdzierz, G.J., Murphy, T.J., & Trimakas, K. (1992). A comparison of non-adjusted and bootstrapped methods: Bootstrapped diagnosis might not be worth the trouble. Educational and Psychological Measurement, 52(1), 181-187.

Lavely, C., Berger, N., Blackman, J., Follman, J., & Kromrey, J. (1992). Content validation of teacher subject matter test ratings of knowledge and skill statements by teachers vs teacher-trainers. Educational and Psychological Measurement, 52(3), 615-622.

McCarroll, D., Crays, N., & Dunlap, W.P. (1992). Sequential ANOVAs and Type I error rates. (1992), Educational and Psychological Measurement, 52(2), 387-393.

Mathews, T.A., & Martin, D.J. (1992). Reciprocal suppression and interaction effects of age with undergraduate grades and gre on graduate performance in a college of education. Educational and Psychological Measurement, 52(2), 453-456.

Mazor, K.M., Clauser, B.E., & Hambleton, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52(2), 443-451.

Roth, W., Gabel, D., Brown, L., & Rice, D. (1992). A combined method of INDSCAL and cognitive mapping for describing changes in students' cognitive structure. Educational and Psychological Measurement, 52(3), 769-779.

van der Ven, A.H.G.S. (1992). Item homogeneity in verbal tests: A Rasch analysis of Amthauer's verbal tests. Educational and Psychological Measurement, 52(3), 623-639.

Wallbrown, F.H., & Jones, J.A. (1992). Reevaluating the factor structure of the revised California Psychological Inventory. Educational and Psychological Measurement, 52(2), 379-386.